Exploratory Data Analysis with Python

PRATHEERTH PADMAN

About Me

Freelance Data Scientist

Content Creator

Pluralsight - https://app.pluralsight.com/profile/author/pratheerth-padman

Udacity - Mentor

LinkedIn - https://www.linkedin.com/in/pratheerthpadman/

Agenda

Agenda

1. What are you expected to know?
2. Setting up your environment
3. Introduction to EDA
4. Univariate and Bivariate Analysis
5. Cleaning the data and Correlation Analysis

What Are You Expected to Know?

What Are You Expected to Know?

- Basic to Intermediate level knowledge of Python and Jupyter Notebooks
- Basic knowledge of Statistics
- Beginner experience with a machine learning project [OPTIONAL]

Setting Up Your Environment

Setting Up Your Environment

- You can find all the files required for the course - dataset, packages, exercises and their solutions at - https://resources.oreilly.com/binderhub/exploratory-data-analysis-with-python
- To follow along and do the exercises with no setup needed - click on the BinderHub service

Introduction to Exploratory Data Analysis

Context for E.D.A

- You are a Machine Learning Engineer or Data Scientist who gets an idea for a project
- You collect, organize and clean the dataset
- Before beginning the modelling process, you need to understand the dataset
- Enter E.D.A

What is Exploratory Data Analysis?

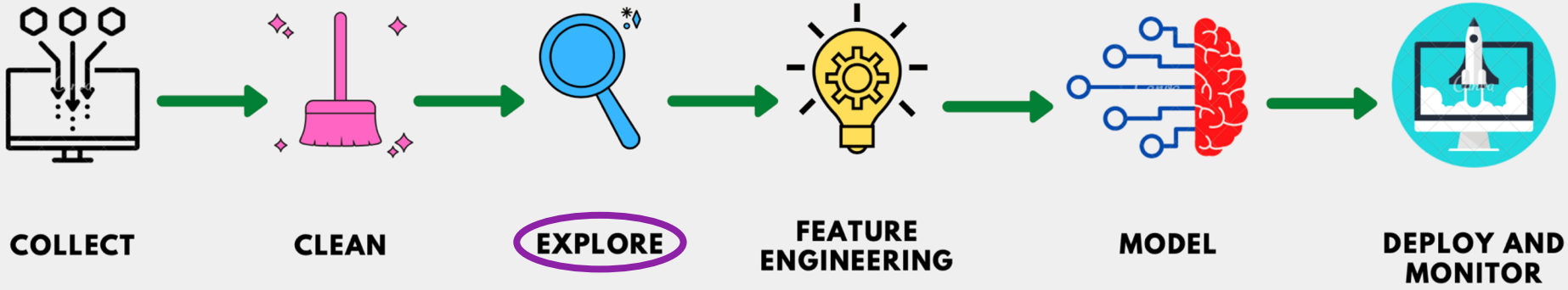
Exploratory Data Analysis is a method of investigating datasets to find preliminary information, insights or uncover underlying patterns in the data.

Instead of making assumptions quickly when seeing the data for the first time, data can be processed in a systematic method to gain insights and make informed decisions.

Importance of E.D.A

- The entire project relies on the data you've collected. Hence, you need to create a solid connection with it.
- You can find errors and discover anomalies in the data. Correcting these if needed will result in better quality data.
- Once you make sense of the data, you can figure out what questions to ask of it. This can lead to brainstorming methods that can better dig into the dataset to find out answers that help solve business or project problems.
- Creation of new features (feature engineering) that help produce better performing machine learning models

Where Does E.D.A Come into the Picture?



Quiz

Q: In a Jupyter Notebook, which among the following commands will display the number of columns in a dataset (dataframe is assigned to “data”)?

- len(data)
- data.shape[1]
- data.shape[0]
- data.head()

Quiz

Q: If a feature in a dataframe contains just whole numbers, what “Dtype” or data type would be assigned to it?

- numerical
- object
- int64
- float64

Univariate & Bivariate Data Analysis

Univariate Data Analysis

Univariate Data Analysis

Univariate data analysis is the simplest form of data analysis where the data being analyzed contains only one variable.

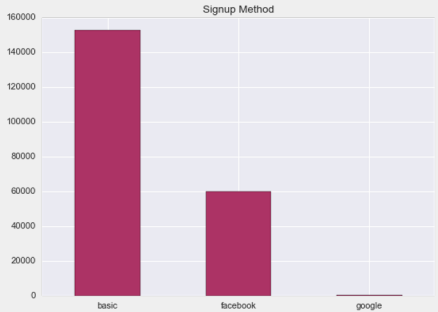
Since it is a single variable, it doesn’t deal with causes or relationships.

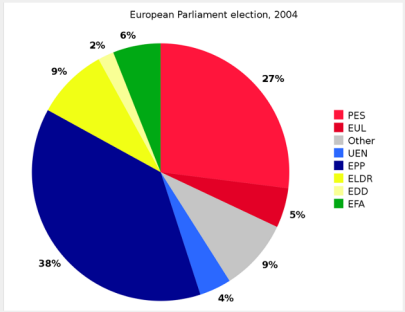
What does Univariate Data Analysis tell us?

- Distribution of data in each feature
- Central tendencies like mean, median and mode
- Max value, min value and standard deviation
- Unique values and missing data

Types of Visualizations - Univariate Data Analysis

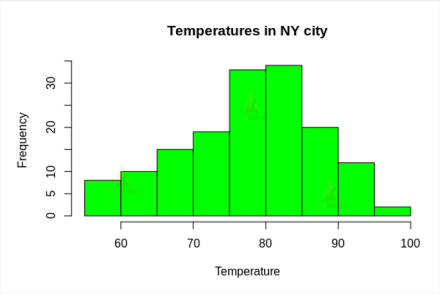
Categorical Feature

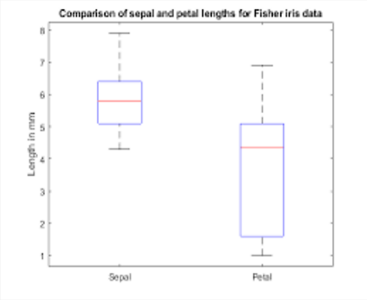


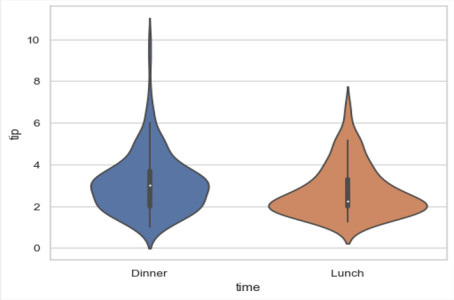


Rank	Degree of agreement	Number
1	Strongly agree	22
2	Agree somewhat	30
3	Not sure	20
4	Disagree somewhat	15
5	Strongly disagree	15

Continuous Feature







Bivariate Data Analysis

Bivariate Data Analysis

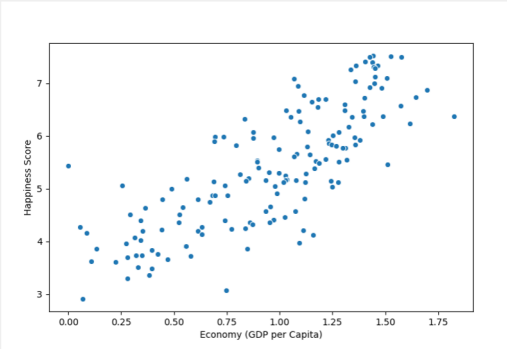
Bivariate data analysis is the analysis of any relationship between two features or variables in a dataset.

What does Bivariate Data Analysis tell us?

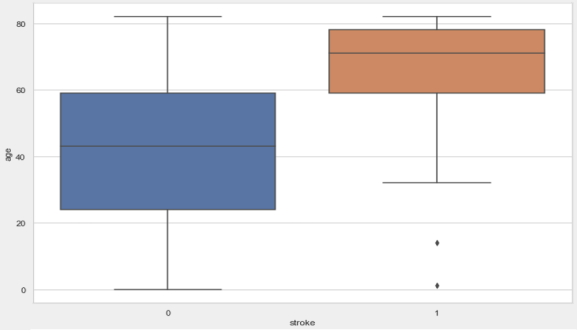
- The existence of a relationship between two variables.
- If one variable influences the other, then we have an independent and dependent variable.
- Positive correlation, negative correlation or no correlation.

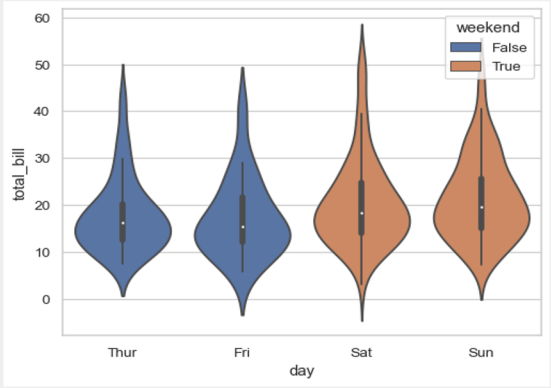
Types of Visualizations - Bivariate Data Analysis

Numerical & Numerical

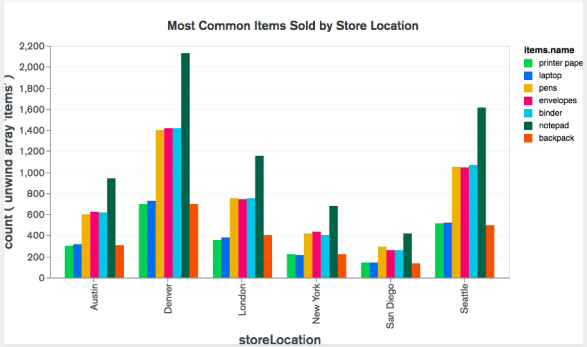


Numerical & Categorical





Categorical & Categorical



Quiz

Q: How does Univariate Analysis differ from Bivariate Analysis?

- Univariate analysis examines one variable, while bivariate analysis looks at the relationship between two variables
- Univariate analysis look at the relationship between two variables, while bivariate analysis examines one variable.
- Univariate analysis is a form of qualitative analysis, while bivariate analysis is a form of quantitative analysis
- Univariate analysis is a form of quantitative analysis, while bivariate analysis is a form of qualitative analysis

Quiz

Q: Barry surveyed his friends to see the number of meals they've eaten from the best restaurant in town in the past month. He received the following responses - 2, 3, 1, 2, 12 and 4. Here, what is "12" an example of?

- Error
- Outlier
- Median
- Mode

Missing Data and Correlation Analysis

Missing Data

What is Missing Data and How Does it Occur?

If values are simply absent or contain NaN (not a number) for any feature in a given dataset, we have missing data. This will cause issues with many machine learning algorithms

How Does it Occur?

- Corrupt data
- Human error during data entry
- Incorrect / Wrong sensor readings
- Software bugs in data processing pipeline

How Do We Deal with Missing Data?

Methods of Dealing with Missing Data

- Remove rows or column with missing data
- Impute with mean or median (numerical feature)
- Impute with mode (categorical feature)
- Imputation with forward fill and backward fill
- Predict the missing values using a machine learning model

Correlation Analysis

Correlation Analysis

Correlation analysis is a statistical method that is used to discover if there is a relationship between two variables or datasets, and how strong that relationship may be.

Correlation Coefficients

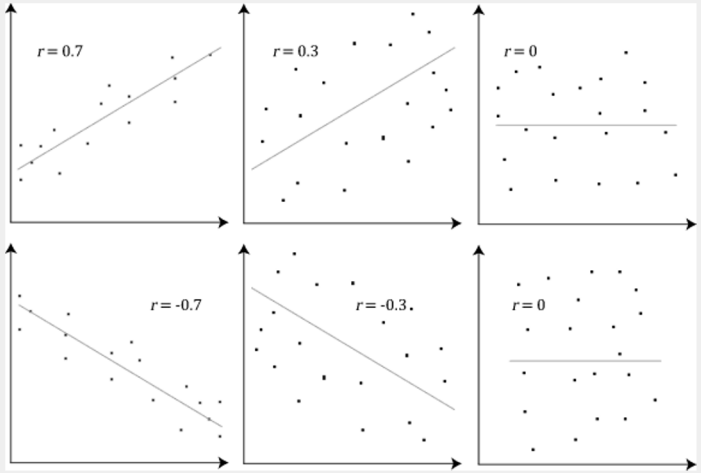
A correlation coefficient is a statistical measure of the strength of the relationship between the relative movements of two variables.

- Spearman Correlation Coefficient
- Pearson Correlation Coefficient

Pearson vs Spearman Correlation Coefficient

Pearson Coefficient

- The Pearson correlation, represented by “r”, can only evaluate a linear relationship between two continuous variables.



Spearman Coefficient

- The spearman correlation, represented by “ρ”, can evaluate a monotonic relationship between two variables and is based on the ranked values for each variable rather than the raw data.

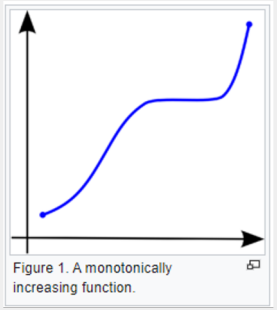


Figure 1. A monotonically increasing function.

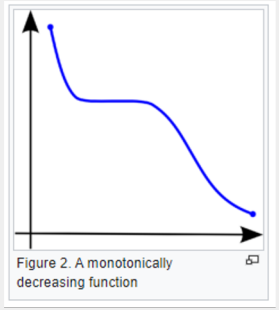


Figure 2. A monotonically decreasing function

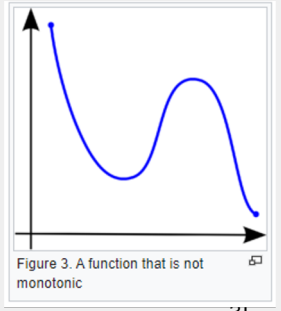


Figure 3. A function that is not monotonic

How Does Correlation Analysis Help?

- Correlation analysis can help to trim the dataset by removing features with very low correlation.

  This would help decrease storage needs and increase model performance.

- We could use this to keep only one of two highly correlated features (multicollinearity).
- Correlation analysis could even help with our choice of model when it comes to machine learning

Correlation and Causation

Correlation techniques indicates if there is any relationship between two variables. **Causation** indicates that one event is the result of the occurrence of the other event, or that there is a causal relationship between the two events.

Causation Example: When a person is exercising then the amount of calories burning goes up every minute. Former is causing the latter to happen.

CORRELATION DOES NOT IMPLY CAUSATION

Correlation Does Not Imply Causation

Sales

Consumption

Even though the sales of sunglasses and consumption of ice cream are highly correlated, it does not mean that one caused the other. It makes no sense to say that since people consume a lot more ice-cream, they also buy a lot of sunglasses. Hence correlation does not imply causation.

If you make decisions based on data, do not jump to conclusions as soon as you find correlation.

Quiz

Q: What will the following code snippet return?

"some_dataframe.isnull()"

- A count of the missing values in each column
- A count of the missing values in each row
- A single boolean value True if the dataframe contains one or more missing values, or False if there are none.
- A corresponding dataframe of boolean values, with True in cells that contain missing values, and False in cells that contain valid data

Quiz

Q: If there were a perfect positive correlation between two variables, the Pearson's r test would give a correlation coefficient of:

- -0.328
- +1
- +0.328
- -1

THANK YOU!